Automated volume analysis of head and neck lesions on CT scans using 3D level set segmentation

Ethan Street, Lubomir Hadjiiski,[a] Berkman Sahiner, Sachin Gujar, Mohannad Ibrahim, Suresh K. Mukherji, and Heang-Ping Chan
Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

(Received 23 January 2007; revised 3 July 2007; accepted for publication 13 September 2007; published 24 October 2007)

The authors have developed a semiautomatic system for segmentation of a diverse set of lesions in head and neck CT scans. The system takes as input an approximate bounding box, and uses a multistage level set to perform the final segmentation. A data set consisting of 69 lesions marked on 33 scans from 23 patients was used to evaluate the performance of the system. The contours from automatic segmentation were compared to both 2D and 3D gold standard contours manually drawn by three experienced radiologists. Three performance metric measures were used for the comparison. In addition, a radiologist provided quality ratings on a 1 to 10 scale for all of the automatic segmentations. For this pilot study, the authors observed that the differences between the automatic and gold standard contours were larger than the interobserver differences. However, the system performed comparably to the radiologists, achieving an average area intersection ratio of 85.4% compared to an average of 91.2% between two radiologists. The average absolute area error was 21.1% compared to 10.8%, and the average 2D distance was 1.38 mm compared to 0.84 mm between the radiologists. In addition, the quality rating data showed that, despite the very lax assumptions made on the lesion characteristics in designing the system, the automatic contours approximated many of the lesions very well. [DOI: 10.1118/1.2794174]

Key words: 3D segmentation, level sets, head and neck cancer, CT scans

I. INTRODUCTION

A vast and rapidly growing body of work in image processing has been devoted to the problem of medical image segmentation. Specifically, an effort is currently underway to design semiautomatic tools for segmentation of lesions in various anatomical regions. One area which has remained relatively nascent in this field is segmentation of head and neck tumors. Currently, three-dimensional segmentations are often obtained via slice-by-slice hand contouring, although various semiautomated tools are beginning to see more common use. Recent work has shown promising progress toward the goal of developing both automatic and semi-automatic tools for producing three-dimensional segmentations.[1–15]

A wide variety of segmentation techniques and models has been developed for the segmentation of anatomical structures and lesions in medical images. A few representative works are summarized in the following. Fuzzy connectedness based segmentation was used by Udupa *et al.*[1] in several medical applications in the areas of multiple sclerosis of the brain, magnetic resonance and CT angiography, brain tumors, upper airway disorders in children, and colonography. Chen *et al.*[2] have applied fuzzy *c*-means-based approach for segmentation of breast lesions in dynamic contrast-enhanced MR images. Active appearance models were used for segmentation of cardiac MR images and diaphragm by Uzumcu *et al.*[3] and Beichel *et al.*,[4] respectively. Another commonly used method is to combine techniques such as seed growing, clustering, and mathematical morphology. Such an approach was used, for example, by Chong *et al.*[5] to semiautomatically segment nasopharyngeal carcinomas.

Deformable contour models, such as active contours and level sets, in particular have seen a recent increase in popularity. Sahiner *et al.*[16] used an active contour model for segmentation of masses in mammograms. Chen *et al.*,[6] Chang *et al.*,[7] and Sahiner *et al.*[8] used active contours for segmentation of breast masses on 3D ultrasound images. Liu *et al.*[9] used a 2D snake algorithm to segment stiff lesions in 3D elastographic ultrasound of *in vitro* tissue specimens.

In several recent studies level sets were used for segmentation of 3D lesions. Cates *et al.*,[12] Droske *et al.*,[13] and Colliot *et al.*[14] applied level sets to brain tumor segmentation. Popovich[15] also used level sets for segmentation of calvarial tumors.

The accuracy of a segmentation method is typically closely related to how specific the task is. A major limitation of many current segmentation methods and especially for the head and neck lesion segmentation is that they are designed under very specific assumptions on the input. Chong *et al.*[5] trained their program by means of user-positioned control points that identified various regions as fat, muscle, bone, etc. Similar work was done by Lee *et al.*,[11] who used clustering and seed growing to segment pixels with different postcontrast enhancement gray level values and ratios of pre- to postcontrast enhancement falling in predefined ranges corresponding to various tissue types. The result is an excellent segmentation but only for a nasopharyngeal carcinoma satis-

fying specific assumptions. Such a procedure could not be expected to work well, for example, on a general class of lesions whose attenuation and size vary depending on location and severity. Complicating the issue is the often striking difference between space-occupying and infiltrating lesions in terms of shape and attenuation.

In this study, we develop a computerized system for automatically segmenting a diverse set of lesions in head and neck CT scans. Our method does not require extensive user interaction or anatomical information. Working under the reasonable assumption that general anatomic abnormalities such as tumors all share common fundamental properties such as having regions of increased gradient as edges, few sharp corners, and an approximately Gaussian distribution in pixel intensity, we have developed a system which can segment a diverse set of lesions in head and neck CT volumes. Our system is based on a 3D level set method, and the only user intervention it requires is to mark a volume of interest (VOI) that encloses the lesion selected for segmentation. In this paper, we describe the methodology used in our segmentation system. We also present preliminary results of our evaluation study in which the computer-generated contours were compared to radiologists' hand-drawn contours via computation of a variety of performance metrics, in addition to subjective ratings of the segmentation quality by a radiologist experienced in interpreting head and neck tumors in CT scans.

II. METHODS

Our computer segmentation system consists of three stages. In the first stage, we apply preprocessing techniques to the original CT images in the 3D volume in order to obtain a set of smoothed images and a set of gradient images. In the second stage, an initial segmentation contour from the preprocessed images is extracted. In the last stage, a serial bank of level sets is propagated from the initial segmentation toward the final segmentation. For clarity of presentation, we will first introduce the level set method and then the preprocessing stages.

II.A. Level sets

The idea behind level sets is to embed a moving contour as the zero set of a time-evolving scalar function $\psi(\mathbf{x})$ defined over the entire image volume.[17,18] This function is interpreted as the signed distance to the contour, and hence the contour is extracted from the function $\psi(\mathbf{x})$ by taking the zero locus. As a convention, points inside the contour have negative values for $\psi(\mathbf{x})$, and those outside have positive values. There are many possibilities in the design of the equation that governs the level set evolution, each with different force terms and coefficients. In this study we have selected a level set with three terms in order to obtain a more complex model with potential to produce accurate 3D segmentation of lesions in head and neck CT volumes, which are diverse in shape and gray level appearance. This type of level set was used also for segmentation of calvarial tumors by Popovich *et al.*[15] Our chosen level set implementation evolves according to the equation

$$\frac{\partial}{\partial t}\psi(\mathbf{x}) = -\alpha\mathbf{A}(\mathbf{x})\cdot\nabla\psi(\mathbf{x}) - \beta P(\mathbf{x})|\nabla\psi(\mathbf{x})| + \gamma\kappa(\mathbf{x})|\nabla\psi(\mathbf{x})|, \quad (1)$$

where the various variables are defined as follows: α, β, and γ are the advection, propagation, and curvature term coefficients, respectively, $\mathbf{A}(\mathbf{x})$ is a vector field image (assigning a vector to each voxel in the image) which causes the contour to move toward regions of high gradient, $P(\mathbf{x})$ is a scalar speed term between 0 and 1 causing the contour to expand at the local rate, and $\kappa(\mathbf{x}) = \mathrm{div}(\nabla\psi(\mathbf{x})/|\nabla\psi(\mathbf{x})|)$ is the mean curvature of the level set at point $\mathbf{x}$.[19–21] The symbol ∇ denotes the gradient operator and "div" is the divergence operator. The purpose of the first two stages of the segmentation system is to generate the images $\mathbf{A}(\mathbf{x})$ and $P(\mathbf{x})$, and also to generate the initial 3D contour from which to begin propagation according to Eq. (1). The purpose of the last stage is to follow the evolution of the level set of $\psi(\mathbf{x})$ over time.

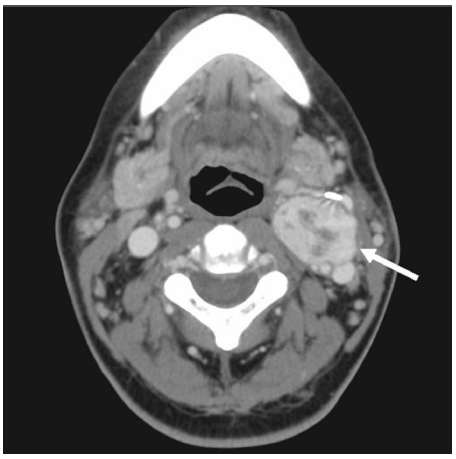

FIG. 1. A CT slice showing a carotid body tumor (difficulty rating of 2) (see white arrow).

II.B. Stage 1: Preprocessing

The goal of the preprocessing step is to generate both a smoothed version of the original image, as well as a suitable gradient image for use in the level set equation. This process is divided into several steps:

II.B.1. Resampling

From the original CT images (an example is shown in Fig. 1), we first crop a box that contains the VOI specified by the radiologist [Fig. 2(a)]. To obtain the cropped volume, we dilate the VOI by 15 pixels (an average of 5.6 mm) on each side in the X and Y directions and two slices (an average of 4.1 mm) in the Z direction on each side. We then linearly interpolate the cropped volume along the Z-axis to produce a 3D raster volume; each slice is interpolated to an integral number of thinner slices such that the z-dimension of a voxel

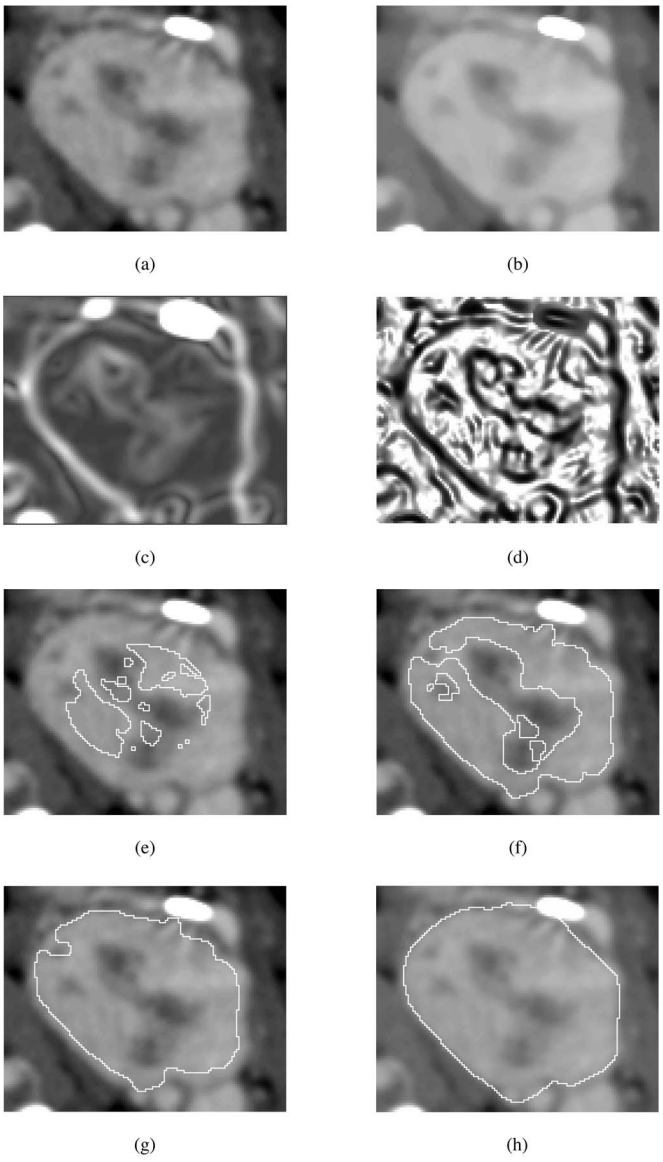

FIG. 2. Segmentation of the carotid body tumor from Fig. 1 by our computer segmentation system: (a) Cropped image. (b) Image after anisotropic diffusion filtering. (c) Gradient magnitude image $|\nabla I(\mathbf{x})|$. (d) Rank transformed image $P(\mathbf{x})$. (e) Boundary of S superimposed on the cropped original image. (f) Boundary of the region $\tilde{C}$ superimposed on the cropped original image. (g) Boundary of the initial segmentation region C superimposed on the cropped original image. (h) Boundary of the final segmentation superimposed on the cropped original image. All operations were performed in 3D. One slice is shown for demonstration.

is closest to those of the x- and y-dimensions. This generally results in an interpolated volume with an interslice spacing to in-slice resolution ratio between 1 and 1.5. Although the voxels are still not isotropic, the differences between the x, y dimensions and the z dimension are reduced while all the information of the original image is maintained.

II.B.2. Smoothing

The image is smoothed using an anisotropic diffusion equation. Anisotropic diffusion acts like Gaussian smoothing but reduces the strength of diffusion near edge pixels, and thus acts as an edge-preserving smoothing filter. If we begin with an image $J(\mathbf{x})$, then the anisotropic equation is a modified version of the heat equation,

$$\frac{\partial J}{\partial t} = \exp\left(-\frac{|\nabla J|^2}{2k^2}\right)\nabla^2 J, \tag{2}$$

where k is a conductance parameter, ∇^2 is the Laplacian, and t is a time variable. The specific parameter values that we used for this filter were $k=2.0$, a time step of 0.05, and 15 iterations. The output of this filter is used as the smoothed image [see Fig. 2(b)], which we denote by $I(\mathbf{x})$.

II.B.3. Normalization

The level set equation requires a speed term $P(\mathbf{x})$ with values between 0 and 1. To obtain this, we first apply a gradient magnitude operator to $I(\mathbf{x})$ [see Fig. 2(c)]. In order to normalize the result, it is necessary to choose a map that transforms the gradient magnitude into the range 0 to 1, with higher gradients mapping to 0. This causes the level set to stop near edges, and to expand in regions of low gradient. Many approaches to the design of the maps have been discussed in the literature. Common choices include the exponential mapping $P(\mathbf{x})=\exp(-\delta|\nabla I(\mathbf{x})|)$, or the inverse mapping $P(\mathbf{x})=(1+\varepsilon|\nabla I(\mathbf{x})|)^{-1}$, where δ and ε are free parameters,[22] and $\nabla I(x)$ is the local gradient at $\mathbf{x}$. Because the best choices for the free parameters depend on the type of images being segmented, these parameters must be tuned to the specific application, which may be a time-consuming task. At the same time, a good quality gradient image is critical because it is the only information the level set has about the underlying image. Instead of applying a contrast change such as the two described above, we use a method known as the "rank transform"[23] defined as follows. We first choose a moving window of user-defined size $(2a+1)\times(2b+1)\times(2c+1)$, where a, b, and c are integral number of pixels in the X, Y, and Z directions, respectively. This window is moved over the image $K(\mathbf{x})=-|\nabla I(\mathbf{x})|$ (the sign of the gradient is inverted so that more negative values correspond to sharper edges), and we count the number of pixels in the window with values less than the center pixel (where we account for a tie by increasing the rank by 1/2). In this way, the center pixel is assigned its rank relative to the rest of the pixels in the window. The rank values are then normalized by the total number of pixels in the window,

$$P(\mathbf{x}) = \frac{\sum_{i=-a}^{a}\sum_{j=-b}^{b}\sum_{k=-c}^{c} u[K(\mathbf{x}) - K(\mathbf{x}+(i,j,k))]}{(2a+1)(2b+1)(2c+1)},$$
$$\text{where } u(t) = \begin{cases} 0, & t<0 \\ \frac{1}{2}, & t=0 \\ 1, & t>0 \end{cases}. \tag{3}$$

The result is an image taking values between 0 and 1. The rank transform has the desirable property that it turns global information, i.e., the distribution in pixel values over a region, into local information at a pixel. In addition, even at a

relatively weak edge, we may still have a significant rank transformed gradient value, which is useful for lesions having some sharp and some subtle edges. For the current application, a window of size $a=b=8$ and $c=3$ was chosen experimentally [see Fig. 2(d)].

II.B.4. Advection image

The last step is to generate the advection image $\mathbf{A}(\mathbf{x})$. This is generated from the speed term $P(\mathbf{x})$ (the rank transformed gradient magnitude) by simply taking the vector gradient: $\mathbf{A}(\mathbf{x})=-\nabla P(\mathbf{x})$, where we use a negative sign so that the vectors point toward regions of high gradient.

Note that the images $\mathbf{A}(\mathbf{x})$ and $P(\mathbf{x})$ are generated in 3D using information in the X, Y, and Z directions. Because we also run 2D level sets, we generate corresponding images $\mathbf{A}_{2D}(\mathbf{x})$ and $P_{2D}(\mathbf{x})$ which use only 2D information in an analogous way.

II.C. Stage 2: Constructing the initial contour

Our prior experience with using an active contour model for segmentation has shown that the final contour is highly sensitive to the initial segmentation.[10] This is complicated by the fact that these models are often not propagated to convergence. In particular, level sets can produce substantially different results for different starting points, especially when the interior of the region being segmented is relatively heterogeneous. For this reason, a major goal of the current research was to develop a method for constructing a high-quality initial segmentation before applying the level set operation. The procedure we have developed appears to give good results for a majority of cases. Several steps are involved as follows.

II.C.1. Sampling

Based on the VOI enclosing the lesion as marked by the radiologist, we define W as the maximal ellipsoid inscribed in the lesion VOI. The ellipsoid is centered at the VOI and has axis lengths the same as the dimensions of the VOI box. Because we make no assumption on the relative pixel intensity of the lesions being segmented, it is necessary to get some estimate of the range of values spanned by the pixels in a particular lesion. Our approach was to first limit the number of pixels in the image for consideration based on attenuation, gradient, and location using some general rules. To do this, we first approximate the lesion center by the ellipsoid $\frac{1}{2}W$ with radii one half of the inscribed ellipsoid W centered in the VOI, forming a binary mask. We then remove the regions of high gradient from $\frac{1}{2}W$ by removing all the pixels $\mathbf{x}$ for which the percentile value of $|\nabla I(\mathbf{x})|$ is in the top 50%. Finally, we remove those pixels which have intensity in the smoothed image below −400 HU in order to eliminate the air volumes located in the throat or nasal cavity. After these procedures we are left with a subset S of pixels which are relatively close to the center of the lesion, and which belong to smooth (low gradient) areas [see Fig. 2(e)].

In our experience, this procedure eliminates most nonlesion pixels from the image. However, the remaining pixels are often only a small subset of the pixels belonging to the full lesion being segmented. Thus, we use this mask as a statistical sample of the full population of pixels in the lesion and we compute the mean μ and standard deviation σ of the pixel values from the smoothed image within the mask, and form a new binary mask as follows.

II.C.2. Thresholding

Let T be the set of all pixels falling within 3.0 standard deviations of the mean of the pixel values in the mask S, within the ellipsoid W, and with values above −400 HU:

$$T=\{\mathbf{x}\in W:|I(\mathbf{x})-\mu|\leq 3.0\sigma,\ \ I(\mathbf{x})>-400\ \text{HU}\}. \qquad (4)$$

The threshold value of −400 HU was selected as a reasonable cutoff point to separate tissue and empty space (air). The HU is −1000 for air, and 0 for water so that −400 HU is close to the midpoint between them. The overall segmentation is not very sensitive to this parameter as long as it is selected between −900 and −100 HU.

We again remove pixels whose gradient magnitude $|\nabla I(\mathbf{x})|$ is in the top 50% of values in the image, to form the tentative initial region $\tilde{C}$ [see Fig. 2(f)]. What is left is often a disconnected subset of the pixels within the VOI. Moreover, as a result of the thresholding process, inhomogeneities in the lesion such as necrosis or fluids may cause the tentative initial region $\tilde{C}$ to have small gaps. To make the system more robust, we implemented a modified flood fill algorithm to form a connected region and to remove holes, as discussed below.

II.C.3. Initial contour

First, a morphological dilation filter is applied to $\tilde{C}$ to dilate it by 2 pixels. A 3D flood fill algorithm is then used to isolate the connected component of the result closest to the image center. The 3D flood fill algorithm is based on a paint bucket operation except that the neighborhood is a 3D cross (six elements in 3D surrounding a seed voxel, cross shape), rather than a 2D cross.[22,24] A morphological erosion filter is then employed to erode the regions by 2 pixels. This process has the effect of connecting nearby components. Finally, we apply a flood fill to the exterior region and invert the result by switching black with white in order to remove holes. The final result is a new binary mask C which is a simply connected subset of the image, roughly approximating the shape of the lesion contained in the VOI [see Fig. 2(g)]. We then pass C to the level set stage of segmentation.

II.D. Stage 3: Level sets

In the final stage of the segmentation program, we apply a bank of level sets in series to the initial contour C, using the images $\mathbf{A}(\mathbf{x})$ and $P(\mathbf{x})$ constructed during the first stage. For each level set, four parameters must be specified: the three scaling coefficients α, β, and γ in Eq. (1), as well as the number of time steps n to run the level set. We first apply

TABLE I. Parameters for the bank of level sets.

Level set:	α	β	γ	n
First	2.0	1.0	1.0	25
Second	5.0	0.45	q	250
Third	1.0	0.0	0.0	15
2D slices	2.0	0.5	0.5	25

three level sets in 3D with a predefined schedule of parameters, and then apply a 2D level set to every slice of the resulting segmentation to get the final contour. Table I lists the parameters used for each level set. The parameter "q" is defined to be a linear function $\sigma M+\phi$ of the 2D diagonal distance M of the VOI box in millimeters (mm), where $\sigma=0.06$, $\phi=-0.11$. Thus, larger VOIs (i.e., larger lesions) will lead to larger γ for the second level set, and the curvature term of which will increase with the VOI diameter. Therefore, the level set will ignore fine inhomogeneities and focus more on the overall lesion shape when segmenting a large lesion. In this way, we can segment large lesions, which tend to be highly heterogeneous.

As seen in Table I, the second level set has the largest number of time steps, and therefore performs most of the refined segmentation. The first level set slightly expands and smoothes the initial contour. This is because the initial contour stage removes regions of higher gradient and thus pushes the contour slightly inside of the true edge, which we regard as the local maximum in the gradient map. The second level set pulls the contour toward the sharp edges, but at the same time it expands slightly in regions of low gradient. The third level set further draws the contour toward sharp edges.

In our experience, working with 3D levels sets has the advantage of producing segmentations in 3D which are cohesive in the Z-direction. This leads to better incorporation of the 3D geometry of the lesion being segmented. However, when the boundary of the 3D surface is viewed on the individual 2D slices, they appear to show an increasingly larger deviation from the object boundary than that obtained by 2D segmentation, as we move from the central slice toward the top and bottom slices of the segmented object, because of the increasingly large angle between the 2D plane and the normal to the 3D surface. Therefore, we take the contour generated by the 3D level sets, and apply to the individual slices a suite of 2D level sets (one on each slice) that have the images $\mathbf{A}_{2D}(\mathbf{x})$ and $P_{2D}(\mathbf{x})$ as input. We only allow the 2D level sets to propagate for a small number of time steps in order to maintain a degree of interslice cohesion. The output of the 2D level sets is taken as the final segmentation [see Fig. 2(h)].

Our segmentation procedure was implemented as a console-based C++ program compiled under the Linux operating system. The program uses the National Library of Medicine's Insight Segmentation and Registration Toolkit (ITK), which is an open source library consisting of implementations for a variety of segmentation and registration algorithms.[22]

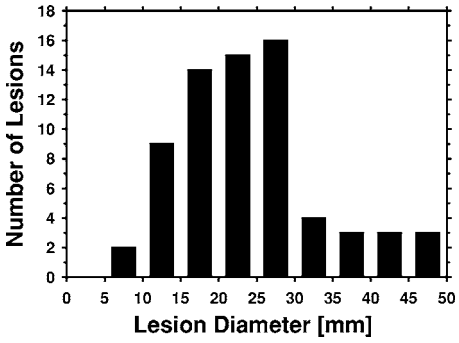

FIG. 3. Distribution of the longest lesion diameter for the data set. The average lesion size was 24.4 mm with standard deviation of 9.4 mm.

III. DATA SET

The segmentation program was evaluated on a preliminary data set consisting of 33 head/neck CT scans from 23 different patients, all collected with IRB approval. Each scan was read by an experienced radiologist who marked the lesions with a bounding box surrounding the lesion and the top and bottom slices using a graphical user interface (GUI) developed in our laboratory. The radiologist also identified the best visualized slice, which was chosen subjectively to be representative of the lesion as a whole and with maximal diameters in the axial planes. The longest diameter and perpendicular diameter on the best slice of each lesion were measured with an electronic ruler. A total of 69 lesions was marked, consisting of primary site carcinomas, lipomas, and lymph nodes. Of the 69 marked lesions, 9 were benign and 60 were malignant. Of the 60 malignant lesions, 25 were biopsy proven to be malignant, while the remaining 35 were determined to be malignant by the experienced radiologist who read the cases and the corresponding clinical reports. The average lesion diameter was 24.4 mm, and the distribution of lesion sizes is shown in Fig. 3. In addition, a difficulty rating scale was defined that represents the radiologist's subjective judgment on the overall conspicuity of a lesion, based on the subtlety of its boundary and overall visibility relative to those encountered in clinical practice. Each lesion was marked with a difficulty rating from 1 to 5, where 1 denotes very obvious and 5 very subtle. Figure 4 plots the distribution of difficulty ratings.

All 2D CT image slices were 512×512 pixels. Scans were acquired with either GE HiSpeed CT/I (single-slice helical), Lightspeed Ultra (8 slice), Lightspeed 16 (16 slice), or Lightspeed Pro 16 (16 slice) scanners. The CT scans used x-ray tube voltages of 120–140 kVp, tube current 140–300 mA, with slice interval of 1.25–3.0 mm and slice thickness of 2.5–3.0 mm. In-plane image resolution varied between 0.351 and 0.585 mm.

In order to test the accuracy of our segmentation system, manually drawn contours were obtained from three radiologists (observers). The contours were drawn using the same

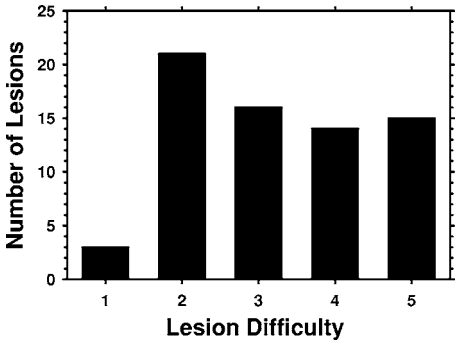

FIG. 4. Distribution of the difficulty ratings for the conspicuity of the lesions in the data set.

GUI mentioned above. Radiologist 1 provided 2D contours for each lesion on the best slice for all 69 lesions. Radiologist 2 provided full 3D contours for lesions 1–12. Radiologist 3 provided the best slice contours for all 69 lesions, and full contours for lesions 13–69. The best slice contours provided by radiologists 1 and 3 were used to define a gold standard set of 2D contours to which the automatically generated segmentations would be compared. To obtain the gold standard set, from each pair of contours of radiologists 1 and 3, we randomly assigned one to set 1, and the other to set 2. This resulted in two sets, with each set containing one contour randomly chosen from the two radiologists for each lesion. Using the GUI, an experienced radiologist chose the best contour out of each pair, and was allowed to modify this contour. The gold standard contours were then selected as the set of contours chosen and modified by the expert. We also combined the 12 3D contours from radiologists 2 and 57 from radiologist 3 to form the gold standard for 3D contours.

IV. EVALUATION METHODS

Several different performance metrics for comparing the similarity of a pair of contours were used in evaluating the system. Although this is not an exhaustive approach, the performance metrics chosen give a reasonably good idea of the degree of similarity between a pair of contours. For each contour U, we are given the information of a polygon on each slice. We first transform the polygon on slice i into a list $L_U^{(i)}$ of adjacent pixel indices in 2D forming a closed curve on a grid. The list $L_U^{(i)}$ therefore represents the set of pixels along the object boundary on slice i. If the contour is three-dimensional, we concatenate the index lists for the adjacent pixel lists on each slice, including the Z-coordinate, yielding a list of points in 3D, denoted by L_U^{3D}. The list L_U^{3D} therefore represents the set of voxels on the object surface. We also form the set of interior points for both the 2D and 3D contours by using a flood fill on every slice, and denote these by $H_U^{(i)}$ and H_U^{3D}, respectively. The performance metrics we used are summarized below.

IV.A. Average dstance

The average distance[25] between two contours U and V is computed as follows. For a fixed point in the boundary list of U, we find the distance to the closest point in the boundary list of V. These distances for all points in U are averaged. We then repeat this process switching the roles of U and V. The two numbers are then averaged,

$$\text{avgdist}(U,V) = \frac{1}{2}\left(\frac{\sum_{x\in A} \min\{d(x,y): y \in B\}}{N_A} + \frac{\sum_{x\in B} \min\{d(x,y): y \in A\}}{N_B} \right), \tag{5}$$

where A denotes either L_U^{3D} or L_U^{best} ("best" denoting the best slice marked by radiologist) and B denotes L_V^{3D} or L_V^{best} for the distance measure in 3D or 2D, respectively; N_A and N_B denote the number of points in A and B, respectively. The function d is the Euclidean distance. Note that this metric is symmetric in the order of its two arguments.

IV.B. Intersection ratio

Given a gold standard contour G and comparison contour U, we compute the intersection ratio as the quotient of the intersection measure and the gold standard measure, in either 2D or 3D, i.e.,

$$R^{2D}(G,U) = \frac{H_G^{best} \cap H_U^{best}}{H_G^{best}}, \tag{6}$$

$$R^{3D}(G,U) = \frac{H_G^{3D} \cap H_U^{3D}}{H_G^{3D}}. \tag{7}$$

A value of 1 implies that contour U completely covers contour G, whereas a value of 0 implies the contours are disjoint.

IV.C. Volume/area error

Given a gold standard contour G and comparison contour U, we compute the volume or area error as the quotient of the difference in volume or area divided by the volume or area of the gold standard, i.e.,

$$E^{2D}(G,U) = \frac{H_U^{best} - H_G^{best}}{H_G^{best}}, \tag{8}$$

$$E^{3D}(G,U) = \frac{H_U^{3D} - H_G^{3D}}{H_G^{3D}}, \tag{9}$$

where positive error indicates oversegmentation and vice versa. From these two performance metrics, volume (area) intersection ratio and volume (area) error, one can derive the true positive fraction, false positive ratio, false negative fraction, and the nonoverlapping volume ratio, as described by Way *et al.*,[10] that provide a complete description of the performance of the segmentation relative to the gold standard. Because the over- and undersegmentation tend to mask the actual deviations from the gold standard when the average is taken, we also reported the absolute (unsigned) errors $|E^{2D}|$

TABLE II. Average and standard deviation of each performance metric for the 2D automatic vs gold standard comparison, the interobserver (2D) comparison, and the 3D automatic vs gold standard comparison.

	2D Measures (best slice)			3D Measures	
Measure		Auto-Gold	Interobserver		Auto-Gold
Intersection Ratio [%]	R^{best}	85.4±12.5	91.2±9.2	R^{3D}	78.3±18.4
Volume or area	E^{best}	4.9±35.8	2.8±16.0	E^{3D}	19.4±59.6
error [%]	$\lvert E^{best}\rvert$	21.1±25.2	10.8±12.7	$\lvert E^{3D}\rvert$	39.6±39.9
avgdist [mm]	avgdist^{2D}	1.38±1.08	0.84±0.68	avgdist^{3D}	1.55±0.87

and $|E^{3D}|$, which estimate the absolute difference between the areas (volumes) of U and G.

IV.D. Quality rating

An additional measure of the automatic segmentation quality is obtained by radiologist's subjective ratings of the contours produced by the system. Each automatic contour was rated by an experienced head and neck radiologist on a scale from 0 to 10 to judge the closeness of the segmentation to the visual lesion boundaries in 3D. The rating scale was designated as follows: 0="unacceptable or missing", 2="very poor", 4="poor", 6="fair", 8="good", 10="excellent or perfect". Specific guidelines were defined for each category, and odd-numbered ratings were defined to be between the adjacent even numbered ratings.

V. RESULTS

The segmentation program was applied to all 69 marked lesions, producing a 3D contour for each. The contours were then compared using the above performance metrics in both 2D (on the slice where the lesion is best visualized) and in 3D. For the 2D measures, we compared the best slice contours extracted by radiologists 1 and 3 in order to establish the interobserver variability. We then compared our automatic contours to the gold standard contours. Because no interobserver results could be obtained for the 3D contours (only one radiologist contoured each lesion in 3D), we simply compared the 3D automatic contour with the corresponding 3D gold standard contour. For evaluation with each performance metric, the average and standard deviation over all 69 lesions were computed. The results are summarized in Table II. For the comparison of automatic versus gold standard, the 2D intersection ratio was $R^{best}=85.4\pm12.5\%$, the absolute area error was $|E^{best}|=21.1\pm25.2\%$, the area error was $E^{best}=4.9\pm35.8\%$, and the average distance measure was $\text{avgdist}^{2D}=1.38\pm1.08$ mm. For the 2D comparison between radiologist 1 and radiologist 3 (the interobserver variation using radiologist 1 as gold standard), the 2D intersection ratio was $R^{best}=91.2\pm9.2\%$, the absolute area error was $|E^{best}|=10.8\pm12.7\%$, the area error was $E^{best}=2.8\pm16.0\%$, and the average distance measure was $\text{avgdist}^{2D}=0.84\pm0.68$ mm. Figure 5 plots the histograms of the results for each measure for the two comparisons. For the 3D measures, the volume intersection ratio was $R^{3D}=78.3\pm18.4\%$, the absolute volume error was $|E^{3D}|=39.6\pm39.9\%$, the volume error $E^{3D}=19.4\pm59.6\%$, and the average distance measure was $\text{avgdist}^{3D}=1.55\pm0.87$ mm. Figure 6 plots the histograms of the results for each measure for the 3D comparison.

Of the 69 segmentations, 40 were given quality ratings of 8 or above. Only 8 were given a rating of under 6 (corresponding to "fair"). The average rating was 7.57, with a standard deviation of 1.65. Figure 7 shows the distribution of lesions with quality ratings.

The average area intersection ratio for the interobserver comparison was approximately 90%. The automatic segmentation achieved an intersection ratio of greater than 90% in over half (38 out of 69) of the lesions when compared to the gold standard. The average interobserver absolute area error was roughly 11%. The automatic system was able to achieve an absolute area error of less than 11% for 28 of the lesions (41%) in comparison with the gold standard.

The average distance comparison had the highest correlation with the radiologist's quality ratings (Pearson's $r=-0.49$). We obtain a significant negative correlation because low average distances imply a close match to the gold

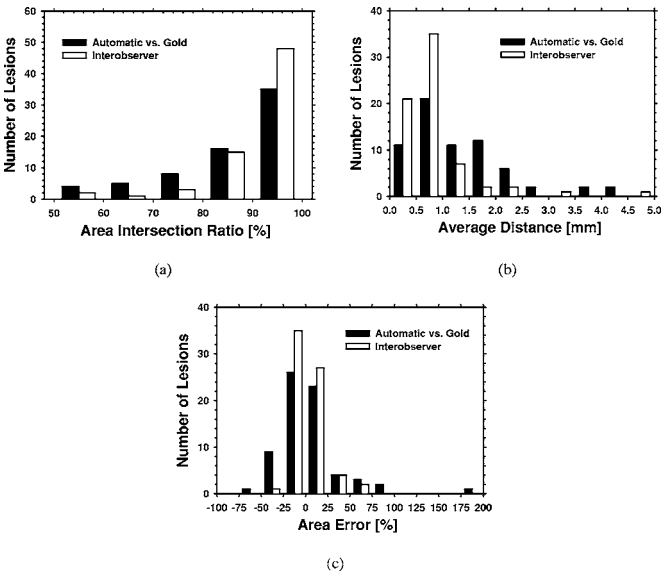

FIG. 5. (a) Histogram of the area intersection ratio measure R^{best}. The average was 85.4% for the automatic vs gold and 91.2% for the interobserver. (b) Histogram of the average distance measure avgdist^{2D}. The average was 1.38 mm for the automatic vs gold and 0.84 mm for the interobserver. (c) Histogram of the area error E^{best} measure. The average was 4.9% for the automatic vs gold and 2.8 % for the interobserver.

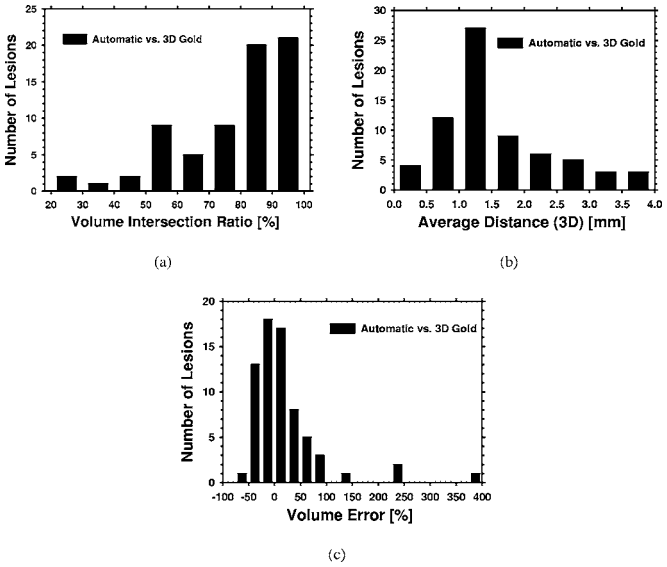

FIG. 6. (a) Histogram of the volume intersection ratio measure R^{3D}. The average was 78.3%. (b) Histogram of the average distance measure avgdist^{3D}. The average was 1.55 mm. (c) Histogram of the volume error E^{3D}. The average was 19.4%.

standard contour and thus a high quality rating. As a result, we regard this performance metric as the best among the three metrics used in this study for predicting the quality of an automatic segmentation. For the 2D comparison, nearly half (32 out of 69) of the lesions segmented had an average distance between the automatic and gold standard contour of less than 1.0 mm, or roughly 2–3 pixels. All but 12 (57 out of 69) had an average distance of less than 2.0 mm, or roughly 5–6 pixels. For the interobserver comparison, 56 had an average distance of less than 1.0 mm between the two radiologist-drawn 2D contours, and 65 had an average distance of less than 2.0 mm.

Of the 69 lesions, 23 had automatic segmentations with area intersection ratio greater than 90% as well as average distance less than 1.0 mm on comparison with the gold standard 2D contours. This includes 13 of the 35 lymph nodes, and 10 of the 32 primary site lesions, so the system does not appear to perform more or less favorably on any specific type of lesions. Further supporting this is the fact that the Pearson's correlation coefficient for the quality rating versus the difficulty rating was only −0.23, implying a very weak relationship.

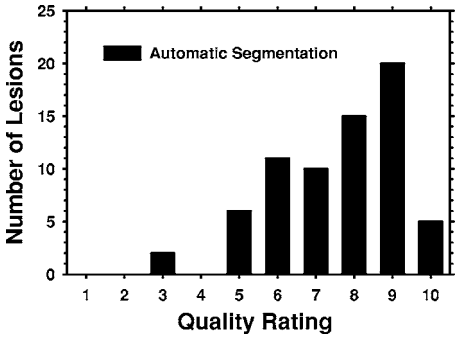

FIG. 7. Histogram for the radiologist's quality ratings of the automatic segmentations.

VI. DISCUSSION

Our segmentation system was designed to segment a diverse set of head and neck lesions given only a bounding box. The goal was to produce a system able to perform reasonably within the realm of interobserver variability. However, a paired t-test applied to the results from the automatic versus gold comparison and the interobserver comparison gives a p-value of less than 0.001 for all three performance metrics, which implies that the deviation of the automatic segmentation from the gold standard is significantly greater than the interobserver variability. In addition, the difference between the 3D automatically generated contours and the 3D radiologist's contours is statistically significant—the p-values of the two-tailed t-test for all four 3D performance metrics (volume intersection ratio, absolute volume error, signed volume error, and average distance) are less than 0.001. However, despite these, our preliminary results show that the system did perform well for a large fraction of the lesions, indicating this is a promising approach that may be used to reduce manual segmentation effort if the system is fully developed.

The choice of the parameters for the preprocessing stages and the level sets was accomplished by extensive experimentation in which each parameter was varied over a reasonable range, and the best parameter within the studied range was chosen based on evaluation of the segmentation results. The initial selection of the parameters was based on logical reasoning for the function and desired level of contribution of every module of the segmentation system. Then, the segmentation system was run with different parameters within a range of the initial selection, until reasonable, refined segmentations were obtained, as indicated by the various performance metrics. This procedure for choosing the parameters was by no means exhaustive, and indeed we leave open the possibility that slightly better choices for the parameters may exist. We also performed a sensitivity analysis based on a number of selected key parameters including the main advection scaling α (level set 2), the main propagation scaling β (level set 2), and the number of iterations n (level set 2). The change in the main advection scaling values had relatively small effect on the segmented contours. The change in the main propagation scaling resulted in the largest change in the segmented contours compared to the other two parameters. A change in the range of 10–54% for main propagation scaling resulted in changes for R^{best} in the range of 0.8–4%, E^{best} 11–56%, avgdist^{2D} 0.2–0.8%, R^{3D} 1–5.3%, E^{3D} 7–37%, and avgdist^{3D} 0.2–1.4%. In all sensitivity experiments the volume and area errors appeared to be very sensitive to changes in the selected parameters. This could be related to the fact that the volume of a sphere changes as the cube of the radius and the area as the square of the radius. Therefore, a relatively small change in the diameter of the lesion could result in larger changes in the volume and the

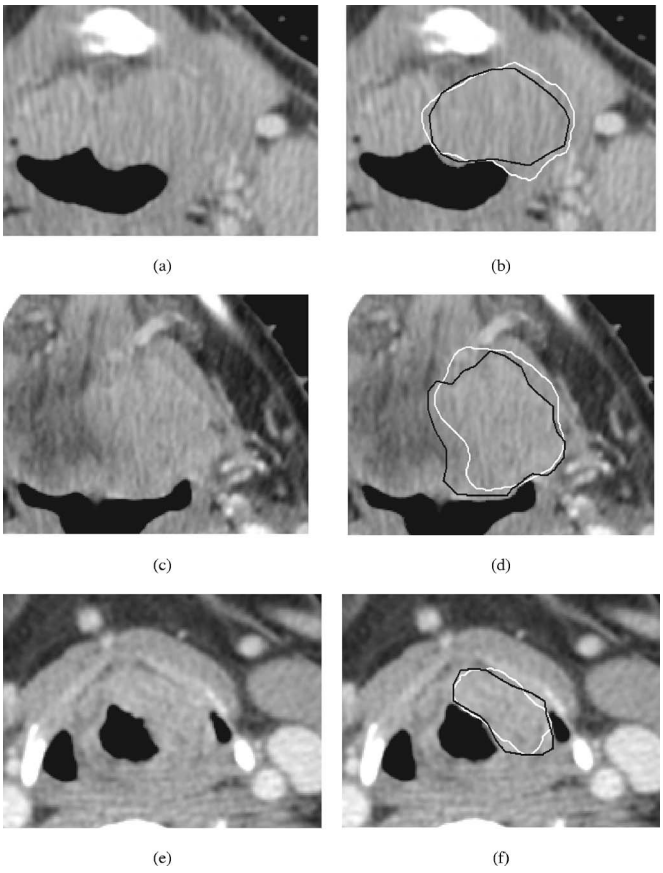

FIG. 8. CT slices of some subtle lesions (difficulty ratings of 4 and 5) in the data set with accurate automatic segmentations. Shown on the left are the original images, and on the right the automatic segmentations (white contours) together with the gold standard (hand-drawn) segmentations (black contours). All lesions are shown on the best slice marked by radiologist. (a), (b) Tongue base carcinoma—difficulty rating of 4. (c), (d) Tongue carcinoma—difficulty rating of 5. (e), (f) Supraglottic carcinoma—difficulty rating of 5.

area of the lesion. On the other hand, the average distance metrics, which was shown to have the highest correlation with the radiologist quality ratings, remained relatively stable and changed very little as these parameters were varied.

The average volume error for the 3D automatic versus gold segmentation comparison was relatively high (19.4%). However, if we analyze the data without the three lesions which have a volume error of over 200%, the average volume error drops to 7.2%, and the standard deviation drops to 37.3% from 59.6%. This indicates that our system in general tends to slightly oversegment on average, but that the degree to which this occurs is exaggerated by the existence of a small number (three) of outliers.

The data set in this study contained a wide range of head and neck lesions of different characteristics. Our segmentation system performs well in some of the lesions visually judged to be most difficult by radiologists, while it fails in others. Figure 8 shows three very subtle lesions (difficulty ratings of 4 and 5), which were accurately segmented by the computer system as compared to the gold standard (hand-drawn) contours. Although most of the boundaries between

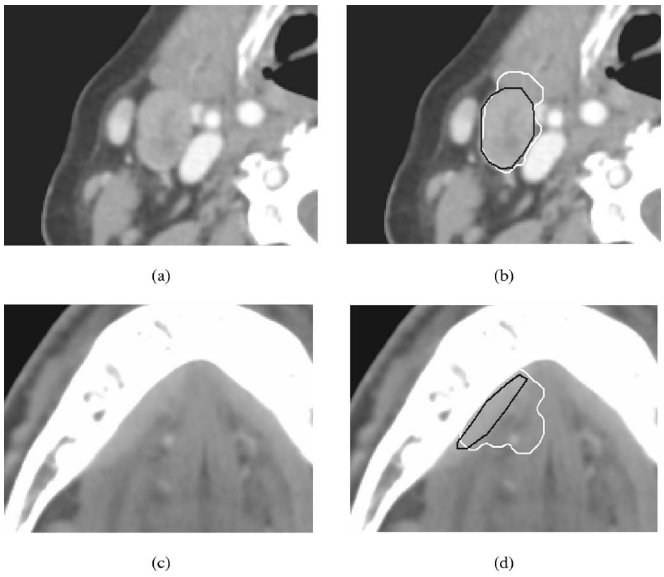

FIG. 9. CT slices of lesions in the data set with poor automatic segmentations. Shown on the left are the original images, and on the right the automatic segmentations (white contours) together with the gold standard (hand-drawn) segmentations (black contours). All lesions are shown on the best slice marked by radiologist. (a), (b) Lymph node—difficulty rating of 2. (c), (d) Floor of mouth carcinoma—difficulty rating of 4.

the lesions and the adjacent normal tissues are very low contrast, the segmentation system was able to estimate reasonable boundaries in these cases. Because the system attempts to estimate the pixel value range for a specific lesion by forming a statistical sample of the pixels near the smooth and central portion of each VOI, the performance of the system can be poor for segmenting highly heterogeneous lesions, such as lymph nodes with necrosis. Figures 9(a) and 9(b) show an example of a lymph node with some mild inhomogeneity in the center that caused the system to mistake the tissue above the lesion, which has similar intensity as the lesion itself, for a part of the object being segmented. In addition, some of the primary site lesions (especially cancers at the floor of the mouth) have very low contrast-to-noise ratio relative to normal tissue such that geometry and pixel intensity alone do not provide enough information to produce a proper segmentation. Figures 9(c) and 9(d) show such an example. The extremely faint lesion edge, in combination with the proximity of lesion to the bright teeth that caused the system to include a broader range of pixel values in the initial contour C, produced an unacceptable segmentation. Anatomical and symmetry considerations, which are not implemented in the current system, may be useful for segmenting these lesions. We will investigate the effects of additional information on the performance of the segmentation system in future studies.

VII. CONCLUSION

The qualitative assessment obtained via the quality ratings showed that, despite the very general assumptions made of the data set when we designed the system, the automatic contours approximated many of the lesions very well. Al-

though the automatic system versus gold standard comparison produced statistically significant differences compared to the interobserver comparison, the results demonstrate that the system performed comparably to the average interobserver variations for a large number of cases. Further investigation is underway to evaluate whether additional information, such as anatomical and symmetry features, could improve the system performance for a greater percentage of lesions.

ACKNOWLEDGMENT

This work is supported in part by USPHS Grant CA93517.

[a]Author to whom correspondence should be addressed. Telephone: (734) 647-7428; Fax: (734) 615-5513. Electronic mail: lhadjisk@umich.edu

[1]J. K. Udupa and P. K. Saha, "Fuzzy connectedness and image segmentation," Proc. IEEE **91**, 1649–1669 (2003).
[2]W. J. Chen, M. L. Giger, and U. Bick, "A fuzzy *c* -means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," Acad. Radiol. **13**, 63–72 (2006).
[3]M. Uzumcu, R. J. V. d. Geest, M. Sonka, H. J. Lamb, J. H. C. Reiber, and B. P. F. Lelieveldt, "Multiview active appearance models for simultaneous segmentation of cardiac 2-and 4-chamber long-axis magnetic resonance images," Invest. Radiol. **40**, 195–203 (2005).
[4]R. Beichel, H. Bischof, F. Leberl, and M. Sonka, "Robust active appearance models and their application to medical image analysis," IEEE Trans. Med. Imaging **24**, 1151–1169 (2005).
[5]V. F. Chong, J. Y. Zhou, J. B. Khoo, J. Huang, and T. K. Lim, "Nasopharyngeal carcinoma tumor volume measurement," Radiology **231**, 914–921 (2004).
[6]D. R. Chen, R. F. Chang, W. J. Wu, W. K. Moon, and W. L. Wu, "3-D breast ultrasound segmentation using active contour model," Ultrasound Med. Biol. **29**, 1017–1026 (2003).
[7]R. F. Chang, W. J. Wu, W. K. Moon, W. M. Chen, W. Lee, and D. R. Chen, "Segmentation of breast tumor in three-dimensional ultrasound images using three-dimensional discrete active contour model," Ultrasound Med. Biol.. **29**, 1571–1581 (2003).
[8]B. Sahiner, H. P. Chan, M. A. Roubidoux, M. A. Helvie, L. M. Hadjiiski, A. Ramachandran, G. L. LeCarpentier, A. Nees, C. Paramagul, and C. Blane, "Computerized characterization of breast masses on 3-D ultrasound volumes," Med. Phys. **31**, 744–754 (2004).
[9]W. Liu, J. A. Zagzebski, T. Varghese, C. R. Dyer, U. Techavipoo, and T. J. Hall, "Segmentation of elastographic images using a coarse-to-fine active contour model," Ultrasound Med. Biol. **32**, 397–408 (2006).
[10]T. W. Way, L. M. Hadjiiski, B. Sahiner, H.-P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," Med. Phys. **33**, 2323–2337 (2006).
[11]F. K. Lee, D. K. Yeung, A. D. King, S. F. Leung, and A. Ahuja, "Segmentation of nasopharyngeal carcinoma (NPC) lesions in MR images," Int. J. Radiat. Oncol. **61**, 608–620 (2005).
[12]J. E. Cates, A. E. Lefohn, and R. T. Whitaker, "GIST: An interactive, GPU-based level set segmentation tool for 3D medical images," Med. Image Anal. **8**, 217–231 (2004).
[13]M. Droske, B. Meyer, M. Rumpf, and C. Schaller, "An adaptive level set method for interactive segmentation of intracranial tumors," Neurol. Res. **27**, 363–370 (2005).
[14]O. Colliot, T. Mansi, N. Bernasconi, V. Naessens, D. Klironomos, and A. Bernasconi, "Segmentation of focal cortical dysplasia lesions on MRI using level set evolution," Neuroimage **32**, 1621–1630 (2006).
[15]A. Popovic, T. Wu, M. Engelhardt, and K. Radermacher, "Modeling of intensity priors for knowledge-based level set algorithm in calvarial tumors segmentation," Lect. Notes Comput. Sci. **4191**, 864–871 (2006).
[16]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," Med. Phys. **28**, 1455–1465 (2001).
[17]R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," IEEE Trans. Pattern Anal. Mach. Intell. **17**, 158–175 (1995).
[18]J. A. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision and Materials Sciences* (Cambridge University Press, Cambridge, 1996).
[19]S. Osher and J. A. Sethian, "Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations," J. Comput. Phys. **79**, 12–49 (1988).
[20]L. Alvarez, P. Lions, and J. Morel, "Image selective smoothing and edge-detection by nonlinear diffusion. 2," SIAM J. Numer. Anal. **29**, 845–866 (1992).
[21]S. Osher and R. P. Fedkiw, "Level set methods: An overview and some recent results," J. Comput. Phys. **169**, 463–502 (2001).
[22]Insight Software Consortium, The ITK Software Guide, 2nd ed., http://www.itk.org (2005).
[23]J. Banks and M. Bennamoun, "Reliability analysis of the rank transform for stereo matching," IEEE Trans. Syst., Man, Cybern., Part B: Cybern. **31**, 870–880 (2001).
[24]M. Fathi and J. Hiltner, "A new fuzzy based flood-fill algorithm for 3D NMR brain segmentation," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (1999), Vol. 4, pp. 881–885.
[25]B. Sahiner, N. Petrick, H. P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on characterization," IEEE Trans. Med. Imaging **20**, 1275–1284 (2001).